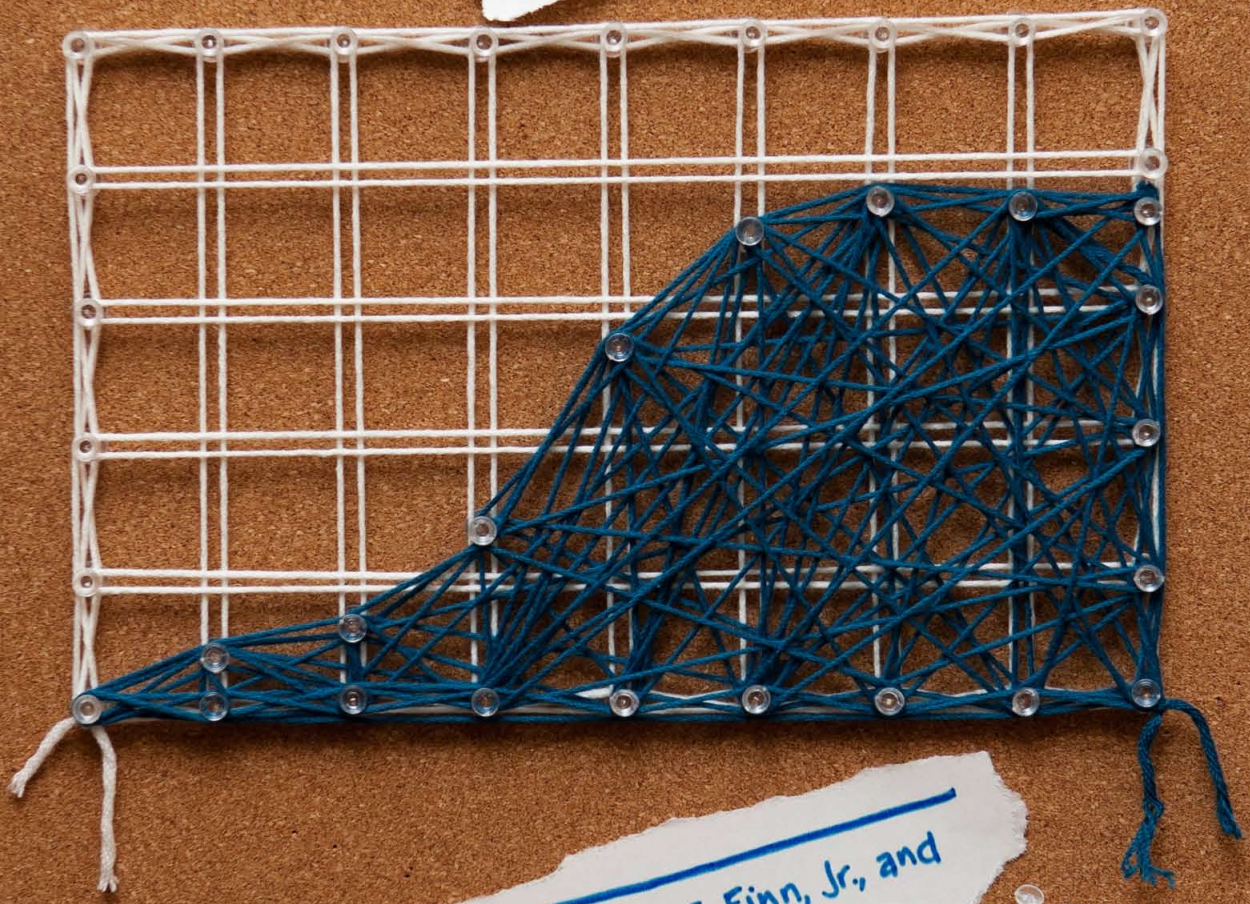
The Accountability Plateau

By Mark Schneider



Foreword by Chester E. Finn, Jr., and Michael J. Petrilli

December 2011

THOMAS B. FORDHAM INSTITUTE
ADVANCING EDUCATIONAL EXCELLENCE

The Accountability Plateau

By Mark Schneider

Foreword by Chester E. Finn, Jr., *and* Michael J. Petrilli

DECEMBER 2011

THOMAS B. FORDHAM INSTITUTE
ADVANCING EDUCATIONAL EXCELLENCE

Contents

Foreword

By Chester E. Finn, Jr., and Michael J. Petrilli

This little paper makes a big point: that "consequential accountability," à la No Child Left Behind and the high-stakes state testing systems that preceded it, corresponded with a significant one-time boost in student achievement, particularly in primary and middle school math. Like the meteor that led to the decline of the dinosaurs and the rise of the mammals, results-based accountability appears to have shocked the education system. But its effect seems to be fading now, as earlier gains are maintained but not built upon. If we are to get another big jump in academic achievement, we're going to need another shock to the system—another meteor from somewhere beyond our familiar solar system.

So argues Mark Schneider, a scholar, analyst, and friend whom we once affectionately (and appropriately) named "Statstud." Schneider, a political scientist, served as commissioner of the National Center for Education Statistics from 2005 to 2008, and is now affiliated with the American Institutes for Research and the American Enterprise Institute. In the following analysis, he digs into twenty years of trends on the National Assessment of Educational Progress—the "Nation's Report Card."

We originally asked Schneider to investigate the achievement record of the great state of Texas. At the time—it feels like just yesterday—Lone Star Governor Rick Perry was riding high in the polls, making an issue of education, and taking flak from Secretary Arne Duncan for running an inadequate school system. We wondered: Was Duncan right to feel "very, very badly" for the children of Texas?[1] Had the state's schools—once darlings of the standards movement and prototypes for NCLB—really slipped into decline since Perry took office? What do the NAEP data really show?

Schneider agreed to take on the project but quickly concluded that there's a larger and more interesting story to tell than simply the saga of Texas. It was true, he noted, that Texas's achievement slowed during the Perry years, particularly as compared to the rest of the country. But rather than pin that development on the governor, Schneider saw a more likely explanation: As an early adopter of standards, testing, and accountability, Texas got a head start on big achievement gains, most of which it realized in the 1990s when George W. Bush was governor—and also a head start on flat-lining thereafter, during Rick Perry's tenure.

Indeed, the Lone Star State made Texas-sized gains from the early- to mid-1990s, as its accountability system got traction. But as other states followed suit, they too hit the achievement fast-track, leading to sizable national gains from 1998 to 2003. Since then, however, Texas's progress has cooled, and the same is now happening to the country as a whole. It's not that Perry was a worse "education governor"

[1] Margaret Talev, "Obama's Education Secretary Says Perry's Schools Left Behind," *Bloomberg*, August 10, 2011, http://www.bloomberg.com/news/2011-08-18/obama-s-education-secretary-says-perry-s-schools-left-behind.html.

than Bush (or, for that matter, Ann Richards) before him, but that he presided over an accountability strategy that was running out of steam.

It's an intriguing argument, and one that deserves serious consideration, even more so as the U.S. marks the tenth anniversary of the enactment of NCLB and tries to figure out what the next version of that law should entail. If school-level accountability, as currently practiced, is no longer an effective lever for raising student achievement, then what is? If we need another "meteor" to disrupt the system, where should we look? Mark suggests that the Common Core and rigorous teacher evaluations have potential. We also see promise in the digital-learning revolution. But other shocks to the system might work even better. What are they?

Introduction

Many educators and elected officials, including more than a few members of Congress, regard "No Child Left Behind," the well-known moniker of George W. Bush's 2001 education act, as a discredited "brand." Indeed, the very acronym NCLB is about to be tossed into the dustbin of history in favor of its progenitor, ESEA (the Elementary and Secondary Education Act), or perhaps some new title yet to be devised on Capitol Hill. There are many reasons why NCLB has been discredited, including, to quote Kevin Carey, the "apocalyptic language out there, that standards and tests have ruined American public education, driven the best teachers out of the classroom, etc., etc."[2]

Yet, as the data presented below demonstrate, NCLB—and the accountability movement it embodied, codified, and symbolized—contributed to a major change in the performance level of American students in math. The data also suggest, however, that the accountability movement has likely reached a point of diminishing (or perhaps even no) returns. While moving on from NCLB is probably essential to produce further growth in student performance, "consequential accountability" was an important and meaningful education reform and ought not be dismissed as a failed initiative.

Debates over the effects and effectiveness of NCLB almost always revolve around national and state scores on the National Assessment of Educational Progress (NAEP). Not surprisingly, the release in November 2011 of the newest NAEP mathematics and reading report cards set off a new round of discussion about the impact of NCLB and accountability more generally. Given the ongoing fights surrounding the overdue reauthorization of ESEA/NCLB, the debate over the effects of accountability is more important now than ever.

Remember that NCLB's system of consequential accountability (in which schools face cascading penalties for failure, e.g., replacement of the school's principal, reconstitution, closure, etc.) was built upon the experience of many states that had already developed such systems before 2001. There is considerable agreement that states adopting consequential accountability before NCLB experienced more rapid growth in their test scores relative to non-adopting states.[3] However, as Hanushek and Raymond note, as NCLB took hold, all states became "effectively consequential accountability states."[4] Perhaps not surprisingly, after NCLB, states that were new to the accountability regime experienced faster growth on NAEP assessments than states that had introduced their own accountability regimes before 2001.

[2] Kevin Carey, "What to Think about the New NAEP Scores," *The Quick and the Ed*, November 1, 2011, http://www.quickanded.com/2011/11/what-to-think-about-the-new-naep-scores.html.

[3] For example, see: Eric A. Hanushek and Margaret E. Raymond, "Early Returns from School Accountability," in *Generational Change: Closing the Test Score Gap*, ed. Paul E. Peterson (Lanham, MD: Rowman & Littlefield Publishers, 2006), 143-166; Martin Carnoy and Susanna Loeb, "Does External Accountability Affect Student Outcomes? A Cross-State Analysis," *Educational Evaluation and Policy Analysis* 24, no. 4 (Winter 2002): 305-331; Thomas Dee and Brian A. Jacob, "Evaluating NCLB," *Education Next* 10, no. 3 (Summer 2010): 54-61; and Tom Loveless, Steve Farkas, and Ann Duffett, *High-Achieving Students in the Era of NCLB* (Washington, D.C.: Thomas B. Fordham Institute, 2008), http://www.edexcellence.net/publications-issues/publications/high-achieving-students-in.html.

[4] Hanushek and Raymond, "Early Returns from School Accountability."

The Case of Texas

Texas was one of the first states in the nation to adopt strict and consequential accountability. The Texas experience was fundamental to the framing of NCLB, as George W. Bush took the lessons and practices of Texas along with him when he moved from Austin to Washington. Thus, examining the growth in NAEP scores in Texas relative to changes in the nation as a whole allows us to tease out some lessons about the effects of accountability on student performance and to speculate about the effectiveness of accountability past, present, and future.

As we look at these data, we should remember that, while NAEP is rightfully viewed as the "gold standard" of assessments, it is not the ideal instrument for detailed statements of cause and effect. We should further keep in mind one of the prime maxims of statistics: Correlation is not causation.

The Remarkable Growth in NAEP Math Scores

It is well known that, as measured by NAEP, American students have improved substantially in math (more in fourth grade than in eighth) and little in reading over the last two decades. Separate and apart from overall averages, there has been continuing concern for the level of skills among racial/ethnic minorities as well as concern for the effects of accountability on low- versus high-performing students (specifically, whether or not NCLB placed so much attention on low-performing students that high-performing students were neglected and suffered as a result). Examining trends in Texas versus the nation presents some insights into these issues.

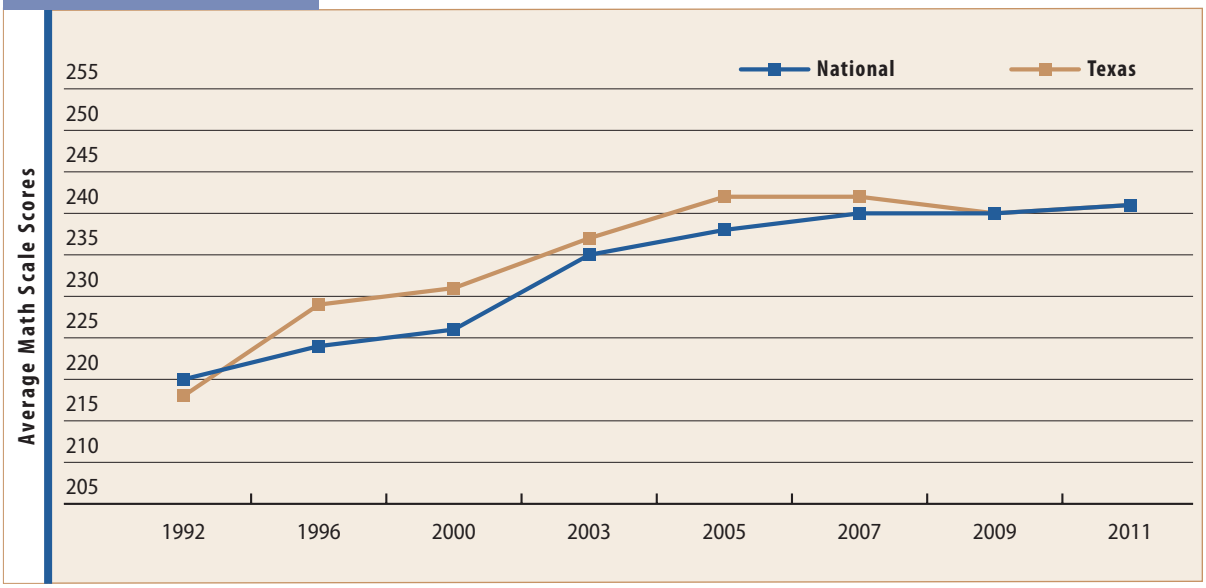
Fourth-Grade Mathematics

Consider Figure 1, which graphs the average scale scores on NAEP's math assessment for fourth-grade students in Texas and in the United States as a whole. The growth in the performance of these students is nothing short of remarkable. Using the very rough rule of thumb that a 10-point change in NAEP scores equals about one year of learning, in 2011 our fourth graders are about two years ahead of where they were in 1992. But, as the figure shows, Texas and the nation marked their peaks of achievement at two distinct points in time.

In 1992, students in Texas were performing at the same level as the students in the nation. In the 1993-94 school year, Texas introduced its system of consequential accountability and, by the time of the next NAEP assessment in 1996, Texas fourth graders had surpassed their peers nationwide. Between 1992 and 2000, math scores across the nation began to creep up; during the same period, a growing number of states began to adopt accountability systems.[5]

[5] Remember that Texas educates about 10 percent of all the nation's students, and national performance is therefore affected by the performance of Texas students. That is, when Texas is outperforming the nation as a whole, the gap would be even larger if we were to exclude Texas from the computation of the national average.

FIGURE 1 **NAEP Fourth-Grade Math (All Students)**



Note: All differences are statistically significant at p<.05 from 1996 to 2007.

By 2003, NCLB had turned every state into a consequential accountability state, and the rate of increase nationwide in math scores between 2000 and 2007 was remarkable. While Texas students continued to outperform the nation as a whole through 2007, the sharp uptick in national performance after 2000 narrowed the Texas lead substantially. Indeed, the last two assessments, in 2009 and 2011, show no significant difference between fourth graders in Texas and fourth graders nationwide.[6]

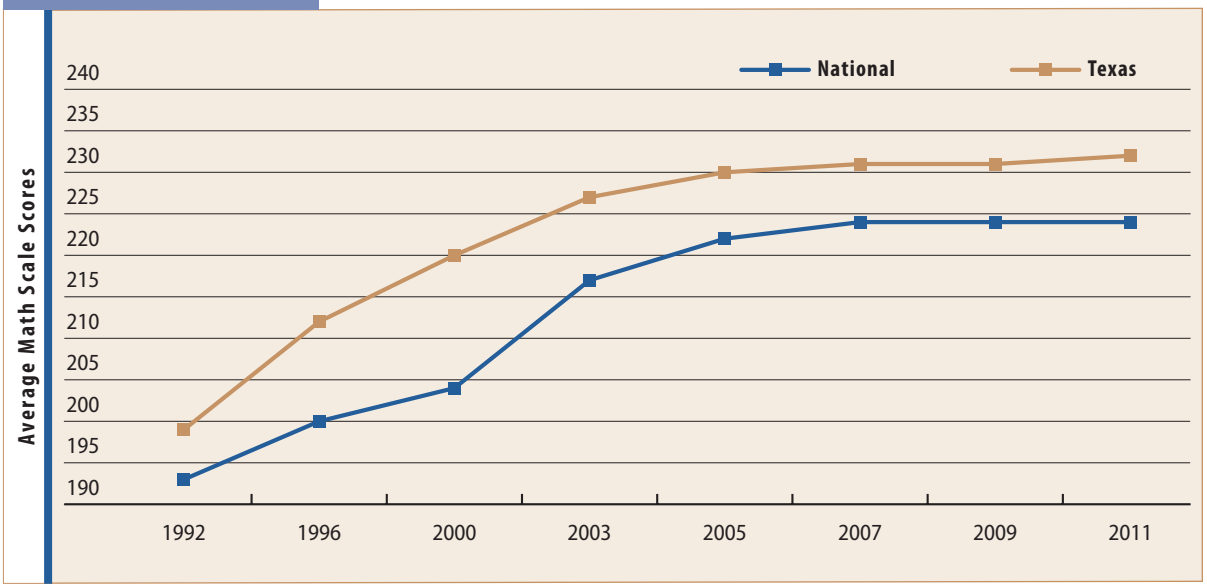
We return to these overall patterns later, but first we turn to the performance of three groups of students who served as particular focal points of NCLB and the accountability movement more generally: blacks (Figure 2), Hispanics (Figure 3), and low-performing students (Figure 4), defined here as those students performing at NAEP's 10th percentile.

At the beginning of the series in 1992, black and Hispanic fourth-grade students in Texas scored slightly higher than their nationwide peers, while those low-performing students at the 10th percentile in Texas achieved at about the same level as those at the 10th percentile nationally.

Between 1992 and 2000, the scores of Texas students in all three groups increased faster than those of their peers nationwide, with the size of the gap between students in Texas and the nation widening to well over 10 points for each group. Between 2000 and 2003, nationwide, the gains for students in each group increased dramatically but then slowed substantially in the years that followed. Gains

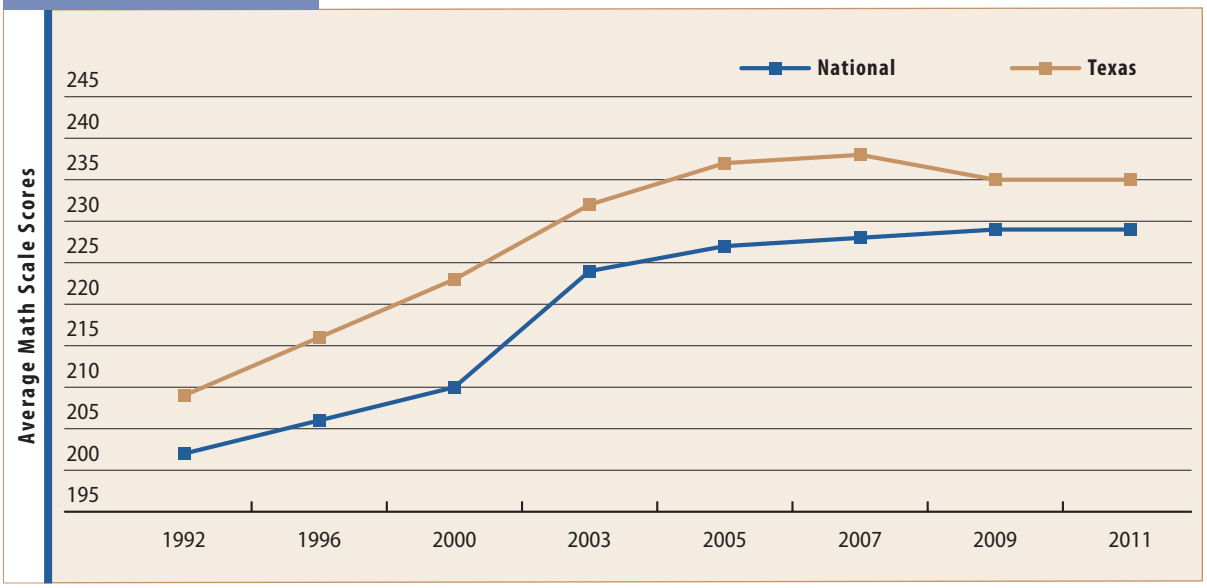
[6] We need to also keep in mind that Texas has far higher-than-expected rates of exclusion on grounds of disability, but the effects are likely not large. For example, between 2007 and 2009, the observed drop in Texas fourth-grade math was 1.9 points. Based on estimates of the full population performance, the drop would have actually been smaller, 1.2 points, but in neither case was the decline significant. See Sami Kitmitto and Victor Bandeira de Mello, *Measuring the Status and Change of NAEP State Inclusion Rates for Students with Disabilities* (Washington, D.C.: National Center for Education Statistics, 2008), http://nces.ed.gov/nationsreportcard/pdf/studies/2009453_3.pdf.

FIGURE 2 NAEP Fourth-Grade Math (Black Students)



Note: All differences are statistically significant.

FIGURE 3 NAEP Fourth-Grade Math (Hispanic Students)

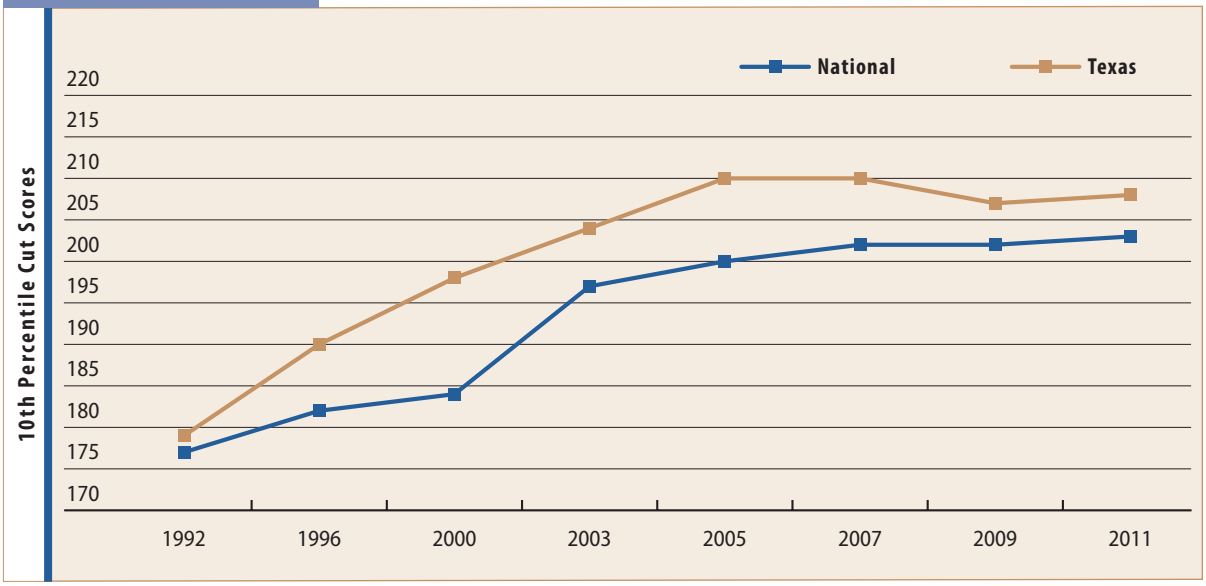


Note: All differences are statistically significant.

among Texas fourth graders were sustained over a longer period of time, but also show evidence of little growth since 2005, with Hispanic and the lowest-performing students actually scoring lower in the latest assessments than in 2007.

The growth in fourth-grade math achievement represents one of the most significant success stories in contemporary American education. Again, the reader is reminded that, while correlation is not causation, the introduction of consequential accountability in Texas and then across the nation coincided

FIGURE 4 NAEP Fourth-Grade Math (Low-Performing Students)



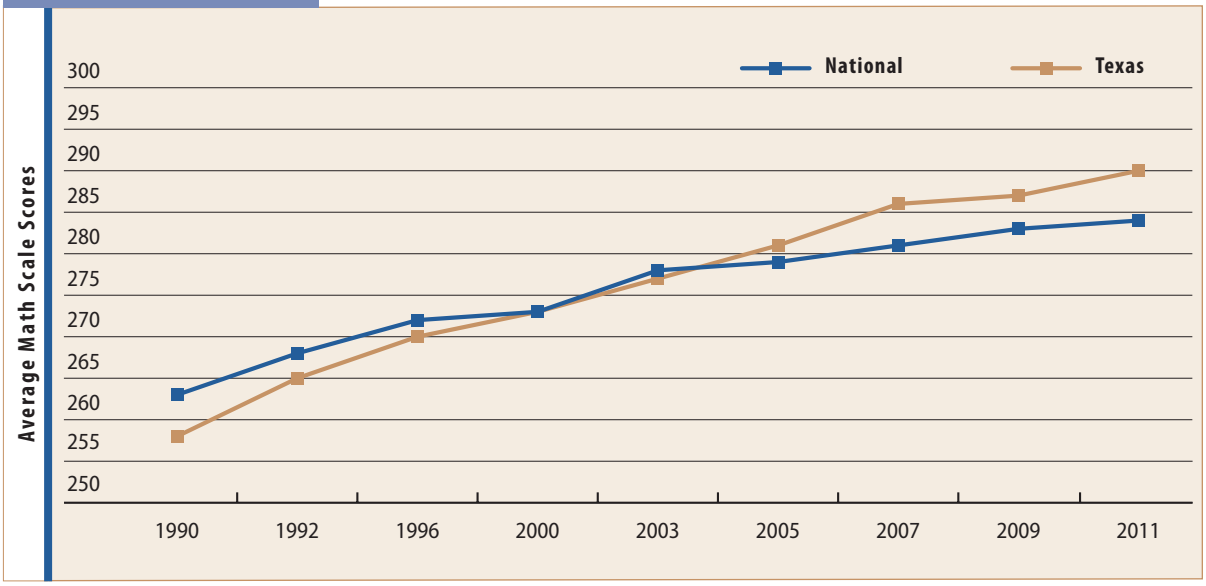
Note: Differences are statistically significant after 1992. Low-performing students are defined as those performing at NAEP's 10th percentile.

with impressive spikes in the performance of students in fourth-grade math, and in particular among the students of most concern to NCLB and the accountability movement more generally.

Eighth-Grade Mathematics

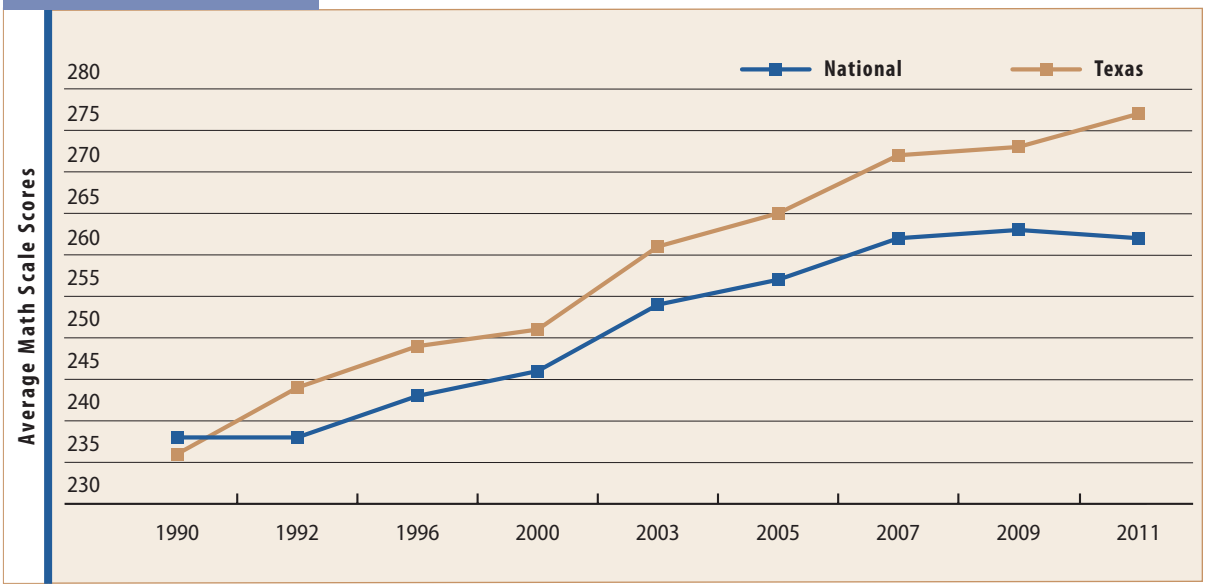
NAEP test results for eighth-grade math represent a somewhat weaker reflection of this striking pattern (Figure 5). The first NAEP eighth-grade math assessment was in 1990, at which time Texas eighth

FIGURE 5 NAEP Eighth-Grade Math (All Students)



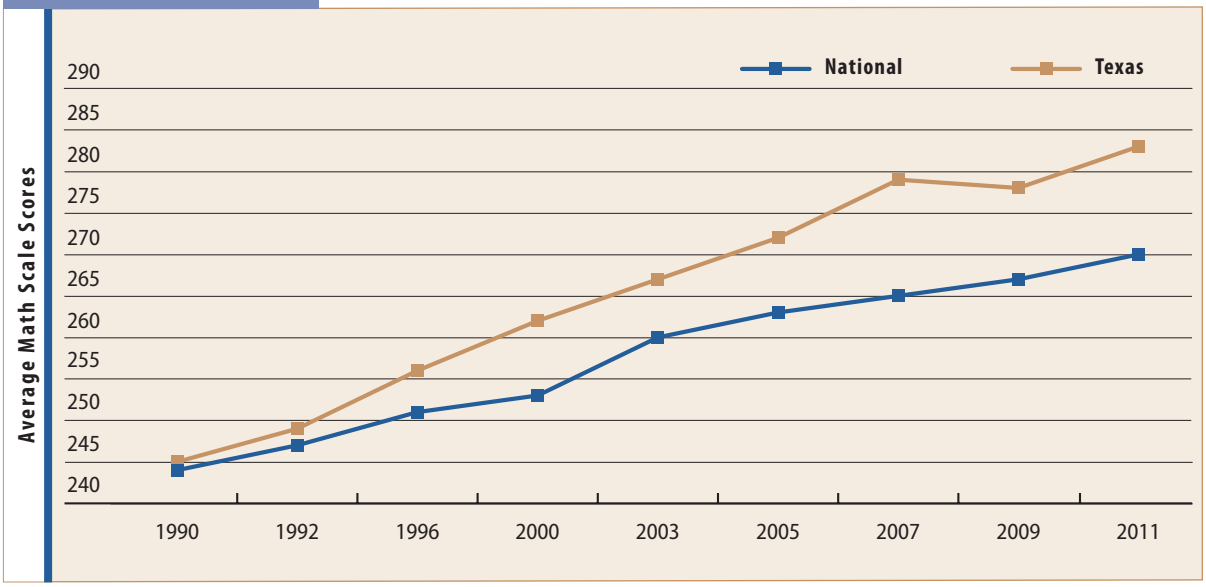
Note: Differences are statistically significant in all years except 2000 and 2003.

FIGURE 6 NAEP Eighth-Grade Math (Black Students)



Note: Differences are statistically significant in all years except 1990 and 2000.

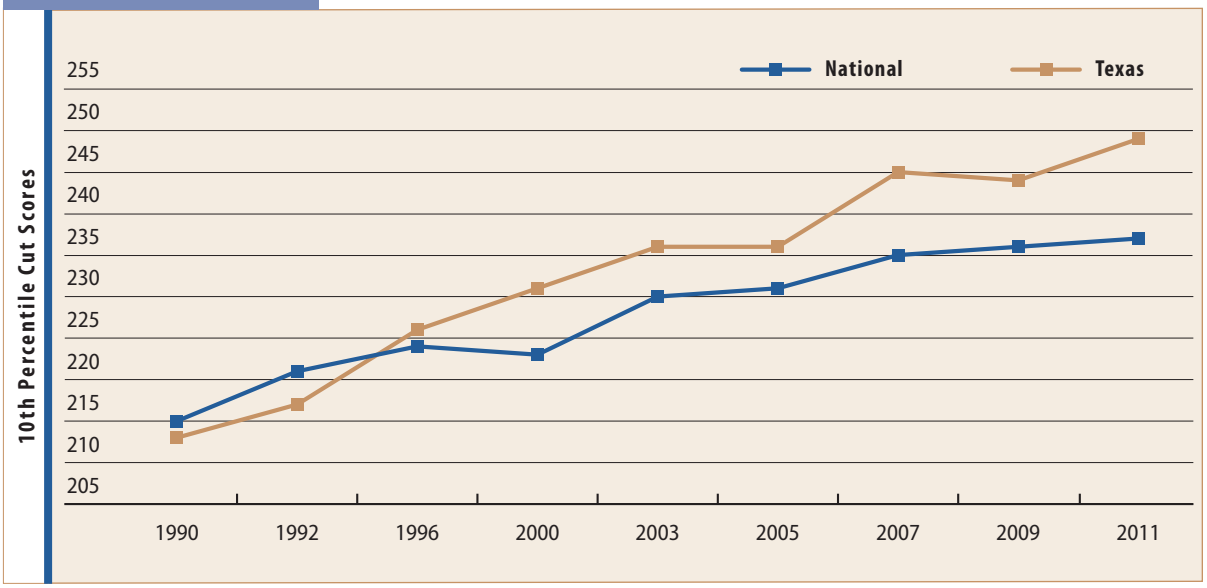
FIGURE 7 NAEP Eighth-Grade Math (Hispanic Students)



Note: Differences are statistically significant from 2000 onward.

graders lagged the nation by 5 points. That gap disappeared by 2000. By 2005, as the strong fourth-grade performers moved into the eighth grade and as the Texas system of consequential accountability continued to gain traction, Texas eighth graders moved past their national peers, producing a gap of 6 points by 2011. Whether eighth-grade test scores can continue to grow, given the flattening scores at the fourth grade, is something that remains to be seen.

FIGURE 8 NAEP Eighth-Grade Math (Low-Performing Students)



Note: Differences are statistically significant from 2000 onward. Low-performing students are defined as those performing at NAEP's 10th percentile.

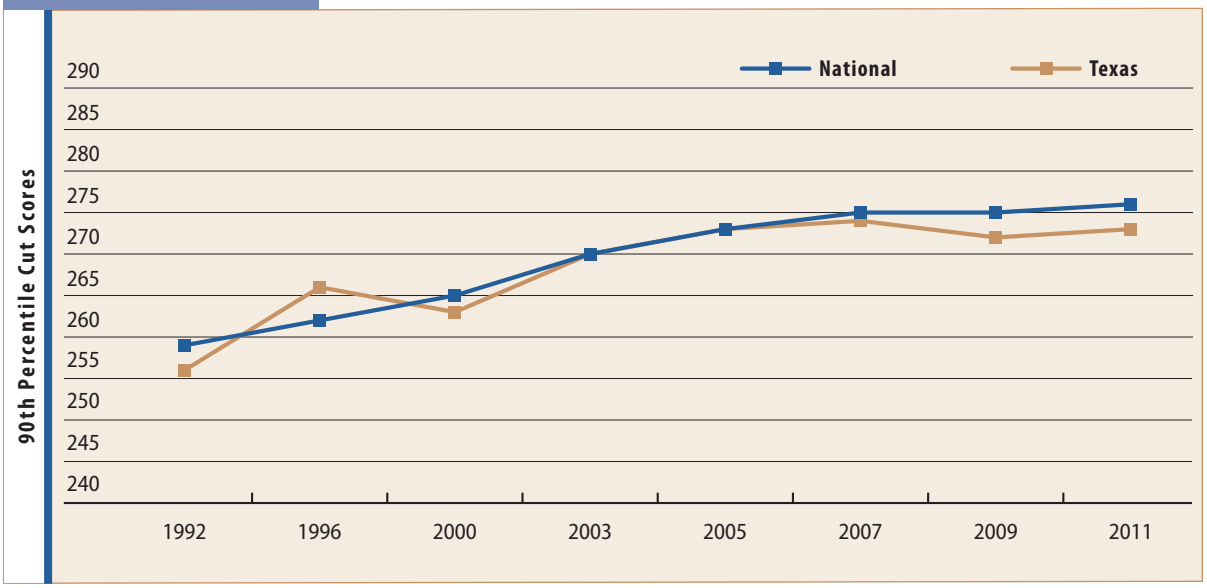
Among black and Hispanic eighth graders, Texas students started at about the same place as their national peers in 1990. Over time, however, they experienced steady growth in performance, producing a widening gap with the nation. Indeed, the size of the gap for black students (in favor of Texas) has increased from 5 or 6 points between 1992-2000 to 10 points or more in the last three assessments (Figure 6). The size of the gaps in favor of Hispanic students in Texas has been somewhat more variable, and was not statistically significant before 2000 (Figure 7). But this gap has grown to over 10 points in the last three assessments. Similarly, the cut score defining the lowest 10th percentile has risen more rapidly in Texas than in the nation as a whole (Figure 8), becoming statistically significant in 2000 and growing from 8 points in 2000 to 12 points in 2011.

High-Performing Students

A frequent criticism of the accountability movement and NCLB was that the focus on racial and ethnic minorities and on the lowest-performing students led to a neglect of the nation's highest-performing youngsters.

Here we define high-performing students as those performing at NAEP's 90th percentile. Fourth-grade math scores for these students both in Texas and in the nation display sharp increases since 1992 (Figure 9). The cut score for the top performers nationwide stood at 259 in 1992 and steadily rose to 276 in 2011, a gain of 17 points. The highest-performing fourth graders in Texas saw a correspondingly large jump in cut scores from 256 in 1992 to 273 in 2005. (Interestingly, half of that gain occurred between the assessments immediately preceding and following implementation of the state's accountability system in 1993-94). Since 2005, however, there has been no statistically significant change in

FIGURE 9 NAEP Fourth-Grade Math (High-Performing Students)

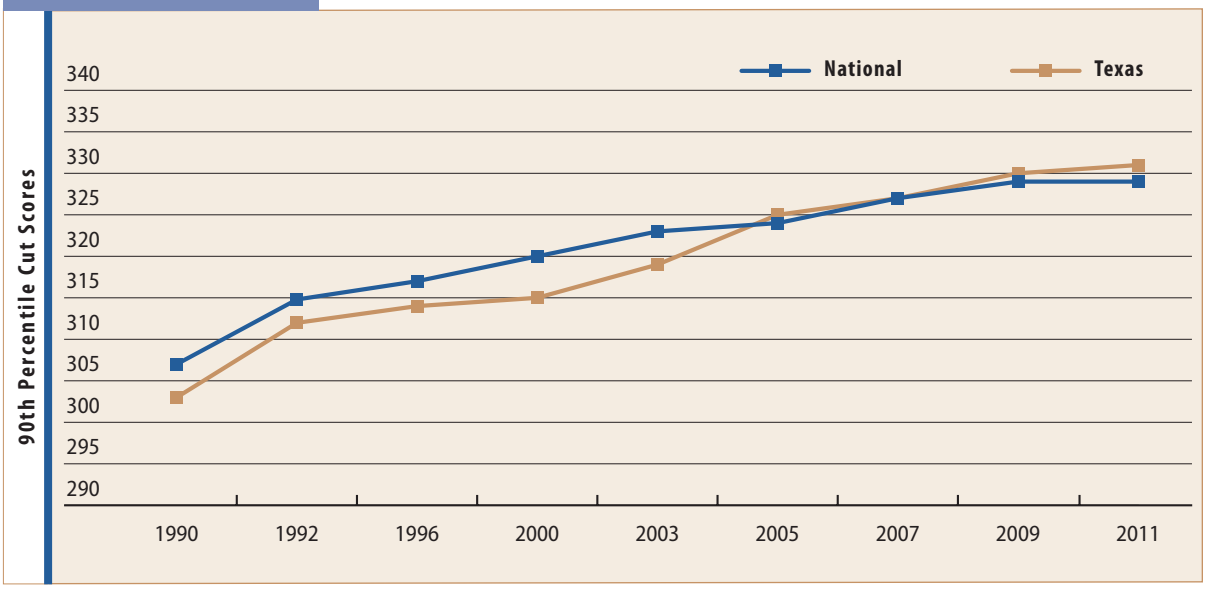


Note: Differences are statistically significant in 2009 and 2011 only. High-performing students are defined as those performing at NAEP's 90th percentile.

cut score for those Texas youngsters, although the national cut score for high performers has continued to rise—producing a statistically significant difference (to the disadvantage of Texas) in the two most recent administrations of NAEP.

Eighth-grade math scores among the highest performers also improved substantially over the period, gaining 14 points nationally and 19 points in Texas from 1992-2011 (Figure 10). The sharpest gains

FIGURE 10 NAEP Eighth-Grade Math (High-Performing Students)



Note: Differences are statistically significant in 2000 and 2003 only. High-performing students are defined as those performing at NAEP's 90th percentile.

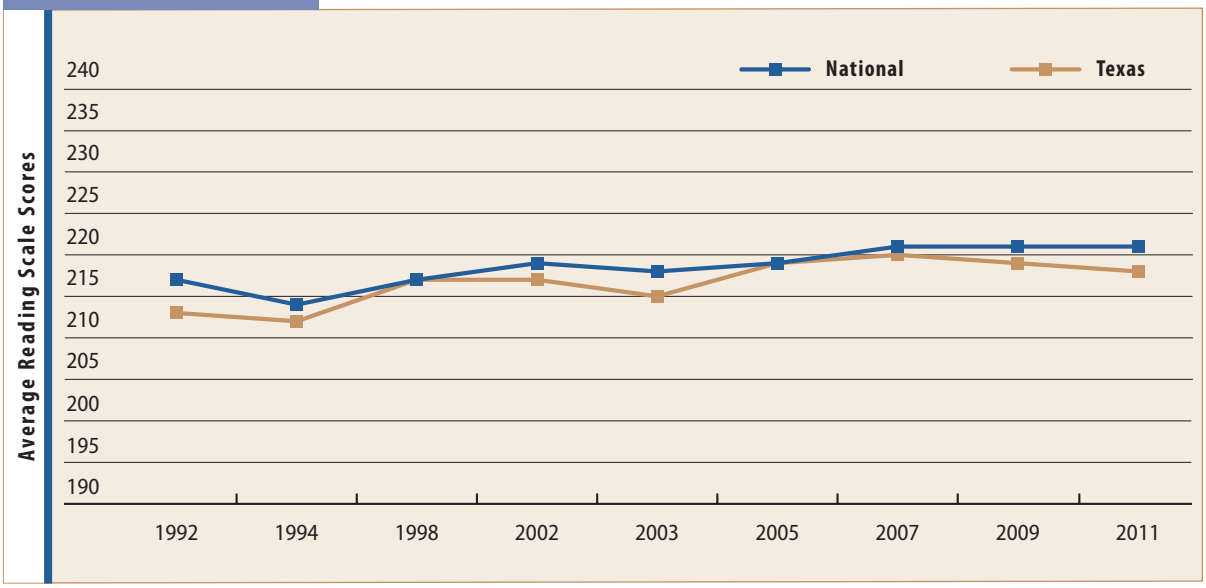
for these high-performing eighth graders in Texas were between 2000 and 2005, building on the improvement made in math by Texas fourth graders four years earlier. Gains continued thereafter at somewhat slower rates, likely reflecting the slower growth in fourth-grade math skills.

The growth in NAEP scores of the highest-performing students in Texas and the nation essentially mirrors the gains made by student groups that were focal to the policy goals of NCLB. Whatever changes more directly focused on specific target populations apparently spilled over to affect the performance of high performers as well. And just as we saw evidence of diminishing effectiveness in recent years for average, minority, and low-performing students, there is evidence that the spillover effects of accountability on high-performing students are also wearing thin. The recent absence of growth in Texas fourth-grade math skills among these high-performing students may portend the end of a remarkable period of growth among the highest performers in the second-largest state in the union.

The Disappointing Case of NAEP Reading Scores

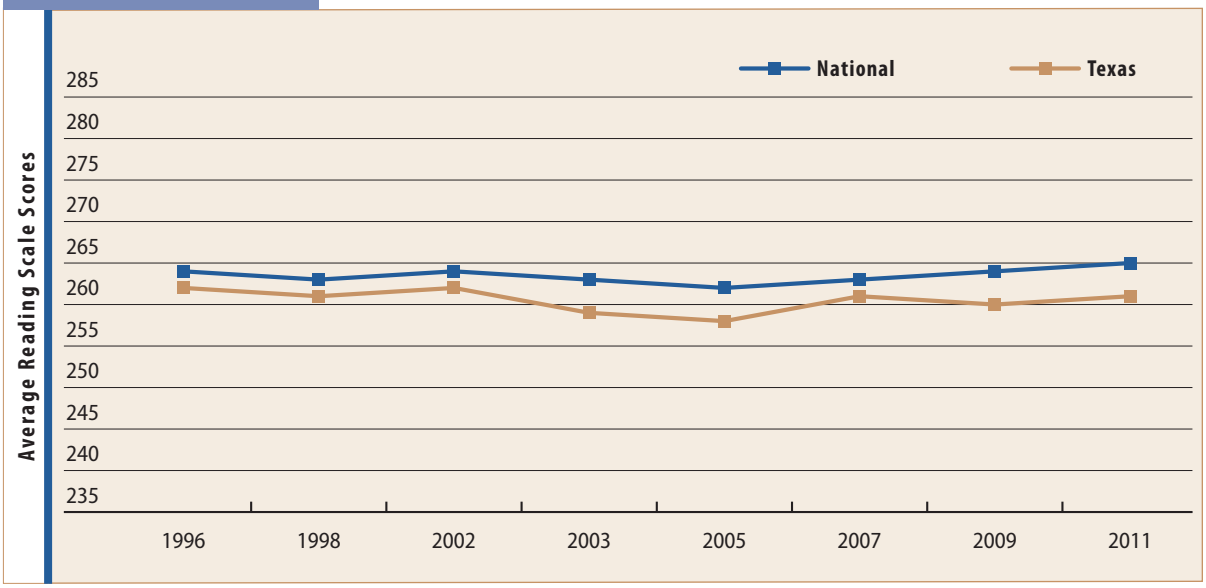
The improvements in NAEP math scores were an unquestionable success for America's fourth and eighth graders and even more so for students in Texas. However, neither the nation as a whole nor Texas has done nearly as well improving students' reading skills. Figure 11 shows no significant difference between the reading scores of fourth-grade students in Texas and in the nation as a whole, except in 2003, and minimal improvement across the board. And Texas's eighth graders have significantly *lagged* the nation since 2003: by 2 points in 2007 and by 4 points in every other assessment between 2003 and 2011 (Figure 12).

FIGURE 11 NAEP Fourth-Grade Reading (All Students)



Note: Difference is statistically significant in 2003 only.

FIGURE 12 **NAEP Eighth-Grade Reading (All Students)**



Note: Differences are statistically significant from 2003 onward.

Accountability and NCLB Were a Success, But…

In 1972, Stephen Jay Gould and Niles Eldredge proposed a theory of evolutionary change that emphasized what they termed "punctuated equilibrium."[7] Their core insight was that complex systems will exist in long periods of stasis. Rather than coming in small incremental steps, change is often characterized by abrupt radical transformations caused by events external to the existing system. Perhaps the most dramatic example is the relatively sudden disappearance of dinosaurs associated with a meteor crashing into the Earth and changing the climate. As a result, the dinosaurs' long reign was replaced by a new equilibrium dominated by mammals.

In 1993, political scientists Frank Baumgartner and Bryan Jones introduced this theory to the study of public policy, and it has since become a common lens through which to view change in social systems.[8] Baumgartner and Jones argued that policy generally changes only incrementally, until some event, such as change in the party control of government or sizable shifts in public opinion, lead to large policy alterations. In their approach, large changes in external conditions (what Baumgartner and Jones term an "exogenous shock") are often needed to produce change in complex social and political systems.

The pattern of test scores in Texas and the nation suggest that consequential accountability—adopted early by Texas, then by more states, and finally by the nation as a whole—was a shock to the U.S. school system that altered the ecosystem and led to a different outcome than had existed before. Over a relatively short period, math performance in fourth and eighth grade abruptly shifted to higher levels of achievement. For example, between 2000 and 2005—the five years spanning the introduction of accountability via NCLB—the average math scale score nationwide at the fourth grade rose by 12 points, roughly a year of learning. In the same period, the average scale score for black fourth graders rose by 18 points, for Hispanic students by 17 points, and the cut score defining the 10th percentile of performance increased by 16 points. The corresponding changes among eighth-grade math scores are small only in comparison: 6 points nationwide, 11 points for black students, 10 points for Hispanic students, and 8 points for those students at the 10th percentile.

To be sure, an important lingering issue is the *absence* of growth in reading scores in Texas and in the nation as a whole. Many have argued that the foundation for reading, compared to math, is far more

[7] Niles Eldredge and Stephen J. Gould, "Punctuated Equilibria: An Alternative to Phyletic Gradualism," in *Models in Paleobiology*, ed. Thomas J. M. Schopf (San Francisco: Freeman, Cooper & Co., 1972), 82-115.

[8] Frank R. Baumgartner and Bryan D. Jones, *Agendas and Instability in American Politics* (Chicago: University of Chicago Press, 1993). Also see Bryan D. Jones, Tracy Sulkin, and Heather A. Larsen, "Policy Punctuations in American Political Institutions," *American Political Science Review* 97, no. 1 (February 2003); Christian Breunig and Chris Koski, "Punctuated Equilibria and Budgets in the American States," *Policy Studies Journal* 34, no. 3 (August 2006); B. Dan Wood and Alesha Doan, "The Politics of Problem Definition: Applying and Testing Threshold Models," *American Journal of Political Science* 47, no. 4 (October 2003).

dependent on what happens early in children's lives—before they enroll in school—and that improving reading skills is therefore much harder to accomplish. Whatever the explanation, clearly the absence of growth reflects a failure of the accountability "meteor" to affect reading levels in a fundamental way.

There is one final pattern to note: As would be expected when viewed through the punctuated- equilibrium lens, once the disruption of consequential accountability has wrung all changes out of the system, a new stasis should take hold. Indeed, Texas, an early adopter, led the nation to higher scores and seems to be ahead of the nation in reaching a new plateau where changes are minimal compared with what came in response to the introduction of an accountability system. The nation, which lagged Texas in adopting accountability, now seems to be entering a period of little change in test scores.

In the 1990s and early 2000s, accountability was an exogenous shock that produced radical gains in math if not in reading. But we now need a new shock to prevent a prolonged period of stasis and stagnation. Scanning the heavens for the next meteor, the most likely candidates to come crashing into the school ecosystem are the Common Core and the better measurement of teacher performance. If the United States is lucky, one or both of these shocks will produce yet another major uptick in math scores. If we are really lucky, these shocks will produce upticks in reading and other subject areas as well.

About the Author

A former commissioner of the U.S. Department of Education's National Center for Education Statistics, Mark Schneider writes about a broad range of education issues: charter schools, consumer choice in education, and higher education policy. He is the author and coauthor of numerous scholarly books and articles, including the award-winning *Choosing Schools: Consumer Choice and the Quality of American Schools* (Princeton University Press, 2000). From 2000 to 2001, he served as vice president of the American Political Science Association (APSA) and simultaneously as president of APSA's public policy section. He is currently vice president at the American Institutes for Research and a visiting scholar at the American Enterprise Institute.

The Thomas B. Fordham Institute is the nation's leader in advancing educational excellence for every child through quality research, analysis, and commentary, as well as on-the-ground action and advocacy in Ohio. It is affiliated with the Thomas B. Fordham Foundation, and this publication is a joint project of the Foundation and the Institute. Many thanks to the Fordham team for assistance on this project, especially Janie Scull for production management and Joe Portnoy and Tyson Eberhardt for dissemination. Erin Montgomery served as copyeditor and Emilia Ryan as layout designer.

For further information, please visit our website at www.edexcellence.net or write to the Institute at 1016 16th St. NW, 8th Floor, Washington, D.C. 20036. The Institute is neither connected with nor sponsored by Fordham University.